Joint estimation of intersecting context tree models

Antonio Galves, Universidade de São Paulo
Aurélien Garivier, CNRS UMR 5141 & Telecom ParisTech
Elisabeth Gassiat, Université Paris Sud CNRS UMR 8628

May 4, 2012

Abstract

We study a problem of model selection for data produced by two different context tree sources. Motivated by linguistic questions, we consider the case where the probabilistic context trees corresponding to the two sources are finite and share many of their contexts. In order to understand the differences between the two sources, it is important to identify which contexts and which transition probabilities are specific to each source. We consider a class of probabilistic context tree models with three types of contexts: those which appear in one, the other, or both sources. We use a BIC penalized maximum likelihood procedure that jointly estimates the two sources. We propose a new algorithm which efficiently computes the estimated context trees. We prove that the procedure is strongly consistent. We also present a simulation study showing the practical advantage of our procedure over a procedure that works separately on each dataset.

Key words: BIC, context tree Models, joint estimation, penalized maximum likelihood, variable length markov chains.

1 Introduction

We assign probabilistic context tree models to data produced by two different sources on the same finite alphabet A. Probabilistic context tree models were first introduced in Rissanen (1983) as a flexible and parsimonious model for data compression. Originally called by Rissanen *finite*

memory source or *probabilistic tree*, this class of models recently became popular in the statistics literature under the name of *Variable Length Markov Chains (VLMC)* Bühlmann & Wyner (1999). The idea behind the notion of variable memory models is that, given the whole past, the conditional distribution of each symbol only depends on a finite part of the past and the length of this relevant portion is a function of the past itself. Following Rissanen we call *context* the minimal relevant part of each past. The set of all contexts satisfies the suffix property which means that no context is a proper suffix of another context. This property allows us to represent the set of all contexts as a rooted labeled tree, by reading the contexts' symbols from the nodes to the root. With this representation, the process is described by the tree of all contexts, called context tree, together with a family of probability measures on A indexed by the contexts. In this work we shall only consider finite context trees. The probability distribution of a context gives the transition probability to the next symbol from any past having this context as a suffix. From now on, the pair composed by the context tree and its family of probability measures will be called *probabilistic context tree*.

The issue we consider here was suggested by a linguistic case study presented in Galves *et al.* (2009). This paper addresses the problem of characterizing rhythmic patterns displayed by two variants of Portuguese: Brazilian and European. This is done by considering two data sets consisting of encoded newspaper texts in two languages. Each data set was analysed separately using a penalized maximum likelihood procedure which selected two different probabilistic context trees corresponding to the two variants of Portuguese. A striking feature emerging from this analysis is the fact that most of the contexts and corresponding transition probabilities are common to the two dialects of Portuguese. Obviously the discriminant features characterizing the different rhythms implemented by the two dialects are expressed by the contexts which appear in one but not in the other model.

To identify those discriminant contexts, the first idea is to estimate separately the context tree for each set of observations, using some classical context tree estimator like the algorithm Context Rissanen (1983) or a penalized maximum likelihood procedure as in Csiszár & Talata (2006) (see also Garivier & Leonardi (2011)), and then compare the obtained trees. This is precisely what is done in Galves *et al.* (2009). However, such an approach does not use the information that the two sources share

some identical contexts and probability distributions. We propose in this paper a selection method using penalized maximum likelihood for the whole set of observations.

In this paper, we argue that a joint model selection more efficiently identifies the relevant features and estimates the parameters. The joint estimation of the two probabilistic context trees is accomplished by a penalized maximum likelihood criterium. Namely, we distinguish two types of contexts: those which appear in both sources with the same probability distribution (we call them *shared contexts*), and the others. The latters appear either in only one of the two sources, or appear in both sources but with different associated probability distributions.

At first sight the huge number of models in the class suggests that such a procedure is intractable. Actually this is not the case. We show that the Context Tree Maximizing procedure, which has been described in Willems *et al.* (1995), can be adapted to recursively find the maximizer: we propose a new algorithm to efficiently compute the estimated context trees. We prove the strong consistency of the procedure. Our proof is inspired by some arguments given in Csiszár & Talata (2006), which handles the case of a single (but possibly infinite) context tree source estimation; as is Garivier (2006), the size of the trees is not bounded in the maximization procedure. We also present a simulation study showing the significant advantage of our procedure, for the estimation of the shared contexts, over a procedure that works separately on each dataset.

The paper is organized as follows. In Section 2, we present the joint context tree estimation problem and the notation. Section 3 is devoted to the presentation of the penalized maximum likelihood estimator we study in this paper. For an appropriate choice of the penalty function, a strong consistency result is given. We describe in Section 4 how to efficiently compute the joint estimator. This is a challenging task, as the number of possible models grows exponentially with the sample size. We show how to take advantage of the recursive tree structure to build an efficient greedy algorithm. The value of this estimator is experimentally shown in Section 5 through a simulation study. The proof of the consistency result is given in Appendix B. It relies on a technical result on the Krichevsky-Trofimov distribution that is given in Appendix A.

2 Notation

Let A be a finite alphabet, and $A^* = \cup_{n\in\mathbb{N}} A^n$ the set of all possible strings including the empty string ϵ. Denote also by $A^+ = \cup_{n\geq 1} A^n$ the set of non-empty strings. A string $s \in A^+$ has *length* $|s| = n$ if $s \in A^n$, and we note $s = s_{1:|s|}$. The empty string has length 0. The *concatenation* of strings s and s' is denoted by ss'. s' is a *suffix* of s if there exists a string u such that $s = us'$; it is a *proper* suffix if $u \neq \epsilon$.

A *tree* τ is a non-empty subset of A^* such that no $s_1 \in \tau$ is a suffix of any other $s_2 \in \tau$. The *depth* of a finite tree τ is defined as

$$D(\tau) = \max\left\{|s| \ : \ s \in \tau\right\} .$$

A tree is *complete* if each node except the leaves has exactly $|A|$ children (here $|A|$ denotes the number of elements in A). Note that $\{\epsilon\}$ is a complete tree.

Let $\mathcal{P}_A$ be the $(|A|-1)$-dimensional simplex, that is the subset of vectors $p = (p_a)_{a\in A}$ in $\mathbb{R}^{|A|}$ such that $p_a \geq 0$, $a \in A$ and $\sum_{a\in A} p_a = 1$. To define a stationary context tree source, we need a complete tree τ and a parameter $\theta \in \mathcal{P}_A^\tau$, that is $\theta = (\theta(s))_{s\in\tau}$ where, for any $s \in \tau$, $\theta(s) \in \mathcal{P}_A$. The A-valued stochastic process $Z = (Z_n)_{n\in\mathbb{Z}}$ is said to be a stationary context-tree source (or variable length Markov Chain) with distribution $\mathbb{P}_{\tau,\theta}$ if for any semi-infinite sequence denoted by $z_{-\infty:0}$, there exists one (and only one) $s \in \tau$ such that s is a suffix of $z_{-\infty:-1}$, and such that, for any $n \geq |s|$, if the event $\{Z_{-n:-1} = z_{-n:-1}\}$ has positive probability, the conditional distribution of Z_0 given $\{Z_{-n:-1} = z_{-n:-1}\}$ is $\theta(s)$ and thus depends only on $z_{-|s|:-1}$. Following Rissanen, an element of τ is called a *context*. In the case when $\tau = \{\epsilon\}$, the source is called *memoryless*.

For any $s \in \tau$, any integer n and any $z_{1:n} \in A^n$, denote by $S(s; z_{1:n})$ the string with the symbols that appear after an occurrence of s in the sequence $z_{1:n}$. Formally,

$$S(s; z_{1:n}) = \bigodot_{i: z_{i-|s|:i-1}=s} z_i \ ,$$

where $\odot$ denotes the concatenation operator. When $z_{i-|s|:i-1} = s$, we say that z_1 is in context s. Besides, denote by $I(z_{1:n}; \tau)$ the set of indices i of $z_{1:n}$ that are not in context s for any $s \in \tau$:

$$I(z_{1:n}; \tau) = \{i \in \{1, \ldots, n\} : \forall s \in \tau, z_{(i-|s|)\vee 1:i-1} \neq s\} \ .$$

Then, if $\mathbb{P}_{\tau,\theta}(Z_{1:n} = z_{1:n}) > 0$,

$$\mathbb{P}_{\tau,\theta}(Z_{1:n} = z_{1:n}) = \prod_{i \in I(z_{1:n};\tau)} \mathbb{P}_{\tau,\theta}(Z_i = z_i | Z_{1:i-1} = z_{1:i-1}) \prod_{s \in \tau} P_{\theta(s)}(S(s; z_{1:n})) \ ,$$

where for $\vartheta \in \mathcal{P}_A$, P_ϑ denotes the probability distribution of the memoryless source on A with parameter ϑ.

Assume that $X = (X_n)_{n\in\mathbb{Z}}$ and $Y = (Y_n)_{n\in\mathbb{Z}}$ are independent stationary context tree sources. Let us define subsets σ_0, σ_1 and σ_2 of A^*, and parameters $\theta_0 = (\theta_0(s))_{s\in\sigma_0}$, $\theta_1 = (\theta_1(s))_{s\in\sigma_1}$, $\theta_2 = (\theta_2(s))_{s\in\sigma_2}$, $\theta_i(s) \in \mathcal{P}_A$, $s \in \sigma_i$, $i = 0, 1, 2$ by the following properties: X has distribution $\mathbb{P}_{\tau_1,(\theta_0,\theta_1)}$, Y has distribution $\mathbb{P}_{\tau_2,(\theta_0,\theta_2)}$, and

$$\sigma_1 \cap \sigma_0 = \emptyset, \ \sigma_2 \cap \sigma_0 = \emptyset, \tag{1}$$

$$\tau_1 := \sigma_1 \cup \sigma_0 \text{ is a complete tree,} \tag{2}$$

$$\tau_2 := \sigma_2 \cup \sigma_0 \text{ is a complete tree,} \tag{3}$$

$$\forall s \in \sigma_1 \cap \sigma_2, \ \theta_1(s) \neq \theta_2(s). \tag{4}$$

σ_0 is the set of shared contexts, that is the set of contexts which intervene in both sources with the same associated probability distributions.

Given two samples $X_{1:n} = (X_1, \ldots, X_n)$ and $Y_{1:m} = (Y_1, \ldots, Y_m)$ generated by X and Y respectively, the aim of this paper is to propose a statistical method for the joint estimation of σ_0, σ_1 and σ_2, and consequently of θ_0, θ_1 and θ_2.

This is a model selection problem, in which the collection of models is described by possible σ_0, σ_1 and σ_2,'s and for fixed σ_0, σ_1 and σ_2 the model consists of all $\mathbb{P}_{\sigma_1\cup\sigma_0,(\theta_0,\theta_1)}$ and $\mathbb{P}_{\sigma_2\cup\sigma_0,(\theta_0,\theta_2)}$ for any possible θ_i, $i = 0, 1, 2$.

We propose in the next section a selection method using penalized maximum likelihood for the entire set of observations.

3 The joint Context Tree Estimator

3.1 Likelihood in context-tree models

For any $(\sigma_0, \sigma_1, \sigma_2)$ satisfying (1), (2) and (3), define $\mathcal{M}_{(\sigma_0,\sigma_1,\sigma_2)}$ as the set of distributions $\mathbb{Q}$ on $A^{\mathbb{N}} \times A^{\mathbb{N}}$ of form

$$\mathbb{Q} = \mathbb{P}_{\sigma_1\cup\sigma_0,(\theta_0,\theta_1)} \otimes \mathbb{P}_{\sigma_2\cup\sigma_0,(\theta_0,\theta_2)} := \mathbb{Q}_X \otimes \mathbb{Q}_Y$$

for some $\theta_0 = (\theta_0(s))_{s\in\sigma_0}$, $\theta_1 = (\theta_1(s))_{s\in\sigma_1}$, $\theta_2 = (\theta_2(s))_{s\in\sigma_2}$, such that $\theta_i(s) \in \mathcal{P}_A$, $s \in \sigma_i$, $i = 0, 1, 2$. Here we do not assume (4).
For any integers n and m, any $x_{1:n} \in A^n$ and $y_{1:m} \in A^m$ and any string s, denote by $S(s; x_{1:n}; y_{1:m}) = S(s; x_{1:n})S(s; y_{1:m})$ the concatenation of the x_i's in context s, and of the y_i's in context s. One has :

$$\begin{aligned} &\mathbb{Q}\left(X_{1:n} = x_{1:n}; Y_{1:m} = y_{1:m}\right) = \\ &\qquad \prod_{i\in I(x_{1:n};\sigma_1\cup\sigma_0)} \mathbb{P}_{\sigma_1\cup\sigma_0,(\theta_0,\theta_1)}\left(X_i = x_i | X_{1:i-1} = x_{1:i-1}\right) \\ &\qquad \prod_{i\in I(y_{1:m};\sigma_2\cup\sigma_0)} \mathbb{P}_{\sigma_2\cup\sigma_0,(\theta_0,\theta_2)}\left(Y_i = y_i | Y_{1:i-1} = y_{1:i-1}\right) \\ &\prod_{s\in\sigma_0} P_{\theta_0(s)}\left(S(s; x_{1:n}; y_{1:m})\right) \prod_{s\in\sigma_1} P_{\theta_1(s)}\left(S(s; x_{1:n})\right) \prod_{s\in\sigma_2} P_{\theta_2(s)}\left(S(s; y_{1:m})\right). \end{aligned} \tag{5}$$

Let us now note for any $s \in A^*$ and any $a \in A$:

$$N_{n,X}(s,a) = \sum_{i=|s|+1}^{n} \mathbb{1}_{X_{i-|s|:i-1}=s, X_i=a},\ N_{n,X}(s) = \sum_{i=|s|+1}^{n} \mathbb{1}_{X_{i-|s|:i-1}=s}$$

where it is understood that an empty sum is 0, and

$$N_{m,Y}(s,a) = \sum_{i=|s|+1}^{m} \mathbb{1}_{Y_{i-|s|:i-1}=s, Y_i=a},\ N_{m,Y}(s) = \sum_{i=|s|+1}^{m} \mathbb{1}_{Y_{i-|s|:i-1}=s}.$$

Observe that $N_{n,X}(\epsilon) = n$ and $N_{m,Y}(\epsilon) = m$. Then, when maximizing over $\mathcal{M}_{(\sigma_0,\sigma_1,\sigma_2)}$ the likelihood as given by (5), we shall use the approximation that the first two terms may be maximized as free parameters (so that their maximization gives 1). Thus we shall use the pseudo maximum log-likelihood

$$\begin{aligned} \ell_{n,m}\Big(\sigma_0, \sigma_1, \sigma_2\Big) &= \sum_{s\in\sigma_1}\sum_{a\in A} N_{n,X}(s,a) \log\left(\frac{N_{n,X}(s,a)}{N_{n,X}(s)}\right) \\ &+ \sum_{s\in\sigma_2}\sum_{a\in A} N_{m,Y}(s,a) \log\left(\frac{N_{m,Y}(s,a)}{N_{m,Y}(s)}\right) \\ &+ \sum_{s\in\sigma_0}\sum_{a\in A} \left[N_{n,X}(s,a) + N_{m,Y}(s,a)\right] \log\left(\frac{N_{n,X}(s,a) + N_{m,Y}(s,a)}{N_{n,X}(s) + N_{m,Y}(s)}\right), \end{aligned}$$

where by convention for any non negative integer p, $0 \log \frac{0}{p} = 0$. Here $\log u$ denotes the logarithm of u in base 2.
For any string s, we shall write $Q_X(\cdot|s)$ and $Q_Y(\cdot|s)$ the probability distributions on A given by: $\forall a \in A$,

$$\begin{aligned} Q_X(a|s) &= \mathbb{Q}\left(X_{|s|+1} = a | X_{1:|s|} = s\right), \\ Q_Y(a|s) &= \mathbb{Q}\left(Y_{|s|+1} = a | Y_{1:|s|} = s\right), \end{aligned}$$

and $\widehat{Q}_X(\cdot|s)$, $\widehat{Q}_Y(\cdot|s)$ and $\widehat{Q}_{XY}(\cdot|s)$ the probability distributions on A given by: $\forall a \in A$

$$\widehat{Q}_X(a|s) = \frac{N_{n,X}(s,a)}{N_{n,X}(s)}\,, \qquad \widehat{Q}_Y(a|s) = \frac{N_{m,Y}(s,a)}{N_{m,Y}(s)}$$
$$\widehat{Q}_{XY}(a|s) = \frac{N_{n,X}(s,a) + N_{m,Y}(s,a)}{N_{n,X}(s) + N_{m,Y}(s)}$$

whenever $N_{n,X}(s) > 0$, $N_{m,Y}(s) > 0$ and $N_{n,X}(s) + N_{m,Y}(s) > 0$ respectively. In the same way, with some abuse of notation, we note Q_X and Q_Y any $|s|$-marginal probability distributions on $A^{|s|}$ defined respectively by $\mathbb{Q}_X$ and $\mathbb{Q}_Y$.

3.2 Definition of the joint estimator

Let pen$(\cdot)$ be a function from $\mathbb{N}$ to $\mathbb{R}$, which will be called penalty function, and define the estimators $\widehat{\sigma}_0$, $\widehat{\sigma}_1$ and $\widehat{\sigma}_2$ as a triple of maximizers of

$$\begin{aligned} C_{n,m}(\sigma_0,\sigma_1,\sigma_2) &= \ell_{n,m}(\sigma_0,\sigma_1,\sigma_2) \\ &\quad - \frac{(|A|-1)}{2}(|\sigma_0|\mathrm{pen}(n+m) + |\sigma_1|\mathrm{pen}(n) + |\sigma_2|\mathrm{pen}(m)) \end{aligned}$$

over all possible $(\sigma_0,\sigma_1,\sigma_2)$ satisfying (1), (2) and (3). The BIC estimator corresponds to the choice pen$(\cdot) = \log(\cdot)$. Notice that it is enough to restrict the maximum over sets $\sigma_0,\sigma_1,\sigma_2$ that have strings s with length $|s| \leq n \vee m - 1$. Indeed, if a string s has length $|s| \geq n$, then for any $a \in A$, $N_{n,X}(s,a) = 0$, if s has length $|s| \geq m$, then for any $a \in A$, $N_{m,Y}(s,a) = 0$.
For any integer D, denote

$$(\widehat{\sigma}_{D,0}, \widehat{\sigma}_{D,1}, \widehat{\sigma}_{D,2}) = \arg\max C_{n,m}(\sigma_0,\sigma_1,\sigma_2)$$

where the maximization is over all $(\sigma_0,\sigma_1,\sigma_2)$ satisfying (1), (2) and (3) and such that for any $s \in \sigma_0 \cup \sigma_1 \cup \sigma_2$, $|s| \leq D$. Then, as explained before, the joint estimator $(\widehat{\sigma}_0, \widehat{\sigma}_1, \widehat{\sigma}_2)$ is seen to be:

$$(\widehat{\sigma}_0, \widehat{\sigma}_1, \widehat{\sigma}_2) = (\widehat{\sigma}_{n\vee m-1,0}, \widehat{\sigma}_{n\vee m-1,1}, \widehat{\sigma}_{n\vee m-1,2})\ .$$

3.3 Consistency of the joint estimator

Now assume that X and Y are independent with distribution

$$\mathbb{Q}^* = \mathbb{P}_{\sigma_1^*\cup\sigma_0^*,(\theta_0^*,\theta_1^*)} \otimes \mathbb{P}_{\sigma_2^*\cup\sigma_0^*,(\theta_0^*,\theta_2^*)}$$

where σ_0^*, σ_1^*, σ_2^* are finite subsets of A^* satisfying (1), (2) and (3), and such that (4) holds.

Theorem 1 *Assume that n and m go to infinity in such a way that*

$$\lim_{n\to\infty} \frac{n}{m} = c,\ 0 < c < +\infty. \tag{6}$$

Assume moreover that for any integer k,

$$\mathrm{pen}\,(k) = \log k.$$

Then the joint estimator is consistent, i.e.

$$(\widehat{\sigma}_0, \widehat{\sigma}_1, \widehat{\sigma}_2) = (\sigma_0^*, \sigma_1^*, \sigma_2^*)$$

$\mathbb{Q}^*$*-eventually almost surely as n goes to infinity.*

We have presented our joint estimator with a generic penalty $\mathrm{pen}(\cdot)$, and Section 4 describes a procedure for computing efficiently this estimator in the general case. However, the consistency result only covers the choice of the BIC penalty Schwarz (1978), that is the penalty which is the logarithm of the number of observations times half the number of free parameters. The proof of Theorem 1 is given in Section B.

4 An Efficient algorithm for the joint estimator

In this section, we propose an efficient algorithm for the computation of the joint estimator with no restriction on the depth of the trees. The recursive tree structure makes it possible to maximize the penalized maximum likelihood criterion without considering all possible models (which are far too numerous). The greedy algorithm we present here can be seen as a non-trivial extension of the Context Tree Maximization algorithm that was first presented in Willems *et al.* (1995), see also Csiszár & Talata (2006). For each possible node s of the estimated tree, the algorithm first computes recursively, from the leaves to the root, indices $\chi_s(X_{1:n}), \chi_s(Y_{1:m})$ and $\chi_s(X_{1:n}; Y_{1:m})$. In a second step, the estimated tree is constructed from the root to the leaves according to these indices.

For any string s let

$$\widehat{P}_s\,(X_{1:n}) = \prod_{a\in A} \left(\frac{N_{n,X}\,(s,a)}{N_{n,X}\,(s)}\right)^{N_{n,X}(s,a)},$$

$$\widehat{P}_s\,(Y_{1:m}) = \prod_{a\in A} \left(\frac{N_{m,Y}\,(s,a)}{N_{m,Y}\,(s)}\right)^{N_{m,Y}(s,a)},$$

and let

$$\widehat{P}_s\left(X_{1:n};Y_{1:m}\right)=\prod_{a\in A}\left(\frac{N_{n,X}\left(s,a\right)+N_{m,Y}\left(s,a\right)}{N_{n,X}\left(s\right)+N_{m,Y}\left(s\right)}\right)^{N_{n,X}(s,a)+N_{m,Y}(s,a)}$$

where again it is understood that for any non negative integer n, $(\frac{0}{n})^0=1$. Notice that, because of possible side effects, $\widehat{P}_s\left(X_{1:n};Y_{1:m}\right)$ is not in general equal to $\widehat{P}_s\left(X_{1:n}Y_{1:m}\right)$.

Step 1: computation of the indices

For any set of strings σ, we denote by σs the set of strings us, $u\in\sigma$: $\sigma s=\{us:u\in\sigma\}$. Let σ be a tree, and let

$$R_{\sigma;s}\left(X_{1:n}\right)=\sum_{u\in\sigma s}\log\widehat{P}_u\left(X_{1:n}\right)-|\sigma|\mathrm{pen}\left(n\right),$$

$$R_{\sigma;s}\left(Y_{1:m}\right)=\sum_{u\in\sigma s}\log\widehat{P}_u\left(Y_{1:m}\right)-|\sigma|\mathrm{pen}\left(m\right),$$

$$R_{\sigma;s}\left(X_{1:n};Y_{1:m}\right)=\sum_{u\in\sigma s}\log\widehat{P}_u\left(X_{1:n};Y_{1:m}\right)-|\sigma|\mathrm{pen}\left(n+m\right).$$

Let D be an upper-bound on the size of the candidate contexts in $\sigma_0\cup\sigma_1\cup\sigma_2$. Note that it is sufficient to consider $D=n\vee m$ to investigate all possible trees. Define for any string of length $|s|=D$:

$$V_s\left(X_{1:n}\right)=R_{\{\epsilon\};s}\left(X_{1:n}\right),\ \chi_s\left(X_{1:n}\right)=0,$$

$$V_s\left(Y_{1:m}\right)=R_{\{\epsilon\};s}\left(Y_{1:m}\right),\ \chi_s\left(Y_{1:m}\right)=0,$$

$$V_s\left(X_{1:n};Y_{1:m}\right)=\max\left\{R_{\{\epsilon\};s}\left(X_{1:n};Y_{1:m}\right);R_{\{\epsilon\};s}\left(X_{1:n}\right)+R_{\{\epsilon\};s}\left(Y_{1:m}\right)\right\},$$

and

$$\chi_s\left(X_{1:n};Y_{1:m}\right)=\begin{cases}1\text{ , if }\ V_s\left(X_{1:n};Y_{1:m}\right)=R_{\{\epsilon\};s}\left(X_{1:n};Y_{1:m}\right)\\ 2\text{ , else.}\end{cases}$$

Then compute recursively for all s such that $|s|<D$:

$$V_s\left(X_{1:n}\right)=\max\left\{R_{\{\epsilon\};s}\left(X_{1:n}\right);\sum_{a\in A}V_{as}\left(X_{1:n}\right)\right\},$$

and

$$\chi_s\left(X_{1:n}\right)=\begin{cases}0\text{ , if }\ V_s\left(X_{1:n}\right)=R_{\{\epsilon\};s}\left(X_{1:n}\right)\\ 1\ \text{ else,}\end{cases}$$

$$V_s\left(Y_{1:m}\right)=\max\left\{R_{\{\epsilon\};s}\left(Y_{1:m}\right);\sum_{a\in A}V_{as}\left(Y_{1:m}\right)\right\},$$

and

$$\chi_s\left(Y_{1:m}\right) = \begin{cases} 0 \text{ , if } V_s\left(Y_{1:m}\right) = R_{\{\epsilon\};s}\left(Y_{1:m}\right) \\ 1 \text{ else.} \end{cases}$$

Define also

$$V_s\left(X_{1:n};Y_{1:m}\right) = \max \begin{cases} R_{\{\epsilon\};s}\left(X_{1:n};Y_{1:m}\right) \\ V_s\left(X_{1:n}\right) + V_s\left(Y_{1:m}\right) \\ \sum_{a\in A} V_{as}\left(X_{1:n};Y_{1:m}\right), \end{cases}$$

and

$$\chi_s\left(X_{1:n};Y_{1:m}\right) = \begin{cases} 1 \text{ , if } V_s\left(X_{1:n};Y_{1:m}\right) = R_{\{\epsilon\};s}\left(X_{1:n};Y_{1:m}\right) \text{ ,} \\ 2 \text{ , if } V_s\left(X_{1:n};Y_{1:m}\right) = V_s\left(X_{1:n}\right) + V_s\left(Y_{1:m}\right) \text{ ,} \\ 3 \text{ else.} \end{cases}$$

For any $(\sigma_0, \sigma_1, \sigma_2)$ satisfying (1), (2) and (3), define

$$R_{(\sigma_1,\sigma_2,\sigma_0);s}\left(X_{1:n};Y_{1:m}\right) = R_{\sigma_1;s}\left(X_{1:n}\right) + R_{\sigma_2;s}\left(Y_{1:m}\right) + R_{\sigma_0;s}\left(X_{1:n};Y_{1:m}\right).$$

Notice that

$$R_{(\sigma_1,\sigma_2,\emptyset);s}\left(X_{1:n};Y_{1:m}\right) = R_{\sigma_1;s}\left(X_{1:n}\right) + R_{\sigma_2;s}\left(Y_{1:m}\right)$$

and

$$R_{(\emptyset,\emptyset,\sigma_0);s}\left(X_{1:n};Y_{1:m}\right) = R_{\sigma_0;s}\left(X_{1:n};Y_{1:m}\right).$$

Moreover, remark that

- either σ_1 and σ_2 are the empty set and σ_0 is not the empty set,
- or σ_0 is the empty set and neither σ_1 nor σ_2 are the empty set,
- or none of them is the empty set.

Step 2: construction of the estimated trees

Once the indicators $\chi_s\left(X_{1:n}\right)$ and $\chi_s\left(X_{1:n}\right)$ have been computed, the estimated sets can be computed recursively from the root to the leaves. Recall that Csiszar and Talata Csiszár & Talata (2006) prove that for any string s such that $|s| \leq D$:

$$V_s\left(X\right) = \max_{\sigma} R_{\sigma;s}\left(X\right) \tag{7}$$

and

$$V_s\left(Y\right) = \max_{\sigma} R_{\sigma;s}\left(Y\right). \tag{8}$$

Call $\sigma_{X_{1:n}}\left(s\right)$ (resp. $\sigma_{Y_{1:m}}\left(s\right)$) a tree maximizing (7) (resp. (8)). $\sigma_{X_{1:n}}\left(s\right)$ and $\sigma_{Y_{1:m}}\left(s\right)$ can be computed recursively as follows: start with the strings s of length D;

- if $\chi_s(X_{1:n}) = 0$, then $\sigma_{X_{1:n}}(s) = \{\epsilon\}$,
- if $\chi_s(X_{1:n}) = 1$, then $\sigma_{X_{1:n}}(s) = \cup_{a\in A}\sigma_{X_{1:n}}(as)\,a$,
- if $\chi_s(Y_{1:m}) = 0$, then $\sigma_{Y_{1:m}}(s) = \{\epsilon\}$,
- if $\chi_s(Y_{1:m}) = 1$, then $\sigma_{Y_{1:m}}(s) = \cup_{a\in A}\sigma_{Y_{1:m}}(as)\,a$.

Namely, for any string s such that $|s| \le D$, define $\sigma_1(s)$, $\sigma_2(s)$ and $\sigma_0(s)$ as:

- if $\chi_s(X_{1:n}; Y_{1:m}) = 1$, then $\sigma_1(s) = \sigma_2(s) = \emptyset$ and $\sigma_0(s) = \{\epsilon\}$,
- if $\chi_s(X_{1:n}; Y_{1:m}) = 2$, then $\sigma_1(s) = \sigma_{X_{1:n}}(s)$, $\sigma_2(s) = \sigma_{Y_{1:m}}(s)$ and $\sigma_0(s) = \emptyset$,
- if $\chi_s(X_{1:n}; Y_{1:m}) = 3$, then $\sigma_1(s) = \cup_{a\in A}\sigma_1(as)\,a$, $\sigma_2(s) = \cup_{a\in A}\sigma_2(as)\,a$ and $\sigma_0(s) = \cup_{a\in A}\sigma_0(as)\,a$.

Validity of the algorithm

The next proposition shows that the two-step procedure described above computes the maximum pseudo-likelihood estimator in the joint model.

Proposition 1 *For any string s such that $|s| \le D$,*

$$V_s(X_{1:n}; Y_{1:m}) = \max R_{(\sigma_1,\sigma_2,\sigma_0);s}(X_{1:n}; Y_{1:m})$$

where the maximum is over all $(\sigma_0, \sigma_1, \sigma_2)$ that verify (1), (2) and (3) and such that

$$\forall u \in \sigma_1 \cup \sigma_2 \cup \sigma_0,\ |u| + |s| = D.$$

In particular,

$$\widehat{\sigma}_{D,0} = \sigma_0(\epsilon)\,,\ \widehat{\sigma}_{D,1} = \sigma_1(\epsilon)\,,\ \widehat{\sigma}_{D,2} = \sigma_2(\epsilon)\,.$$

Proof:

The proof is by induction. Observe first that

$$V_s(X_{1:n}) + V_s(Y_{1:m}) = \max_{\sigma_1,\sigma_2} R_{(\sigma_1,\sigma_2,\emptyset);s}(X_{1:n}; Y_{1:m})\,.$$

Now, if $|s| = D$, then either $\sigma_1 = \sigma_2 = \{\epsilon\}$ and $\sigma_0 = \emptyset$, or $\sigma_1 = \sigma_2 = \emptyset$ and $\sigma_0 = \{\epsilon\}$, and we have

$$V_s(X_{1:n}; Y_{1:m}) = \max\left\{R_{(\{\epsilon\},\{\epsilon\},\emptyset);s}(X_{1:n}; Y_{1:m})\,;\, R_{(\emptyset,\emptyset,\{\epsilon\});s}(X_{1:n}; Y_{1:m})\right\}.$$

Let us now take $|s| < D$ and assume that Proposition 1 is true for all strings as, $a \in A$. The maximum of the $R_{(\sigma_1,\sigma_2,\sigma_0);s}(X_{1:n}; Y_{1:m})$ over all $(\sigma_0, \sigma_1, \sigma_2)$ that verify (1), (2) and (3) and such that $\forall u \in \sigma_1 \cup \sigma_2 \cup \sigma_0,\ |u| + |s| = D$, is reached by a triple $(\sigma_1, \sigma_2, \sigma_0)$ such that:

- either $\sigma_0 = \{\epsilon\}$, in which case σ_1 and σ_2 are necessarily empty and

  $$R_{(\sigma_1,\sigma_2,\sigma_0);s}\left(X_{1:n};Y_{1:m}\right) = R_{(\emptyset,\emptyset,\{\epsilon\});s}\left(X_{1:n};Y_{1:m}\right) = R_{\{\epsilon\};s}\left(X_{1:n};Y_{1:m}\right);$$

- or at least one among σ_1 and σ_2 is equal to $\{\epsilon\}$: then $\sigma_0 = \emptyset$ and

  $$R_{(\sigma_1,\sigma_2,\sigma_0);s}\left(X_{1:n};Y_{1:m}\right) = R_{\sigma_1;s}(X_{1:n})+R_{\sigma_2;s}(Y_{1:m}) = V_s(X_{1:n})+V_s(Y_{1:m})$$

  as in Csiszár & Talata (2006);

- or $\sigma_1, \sigma_2, \sigma_0$ are all different from $\{\epsilon\}$, and then each $\sigma_i, 0 \leq i \leq 2$ can be written as $\sigma_i = \cup_{a\in A}\sigma_i(a)a$; note that it is possible that, for some $i \in \{0,1,2\}$ and some $a \in A$, $\sigma_i(a)$ is empty, or even that σ_i is empty. In any case, for each $a \in A$ it is easily checked that $\sigma_1(a), \sigma_2(a)$ and $\sigma_0(a)$ satisfy (1), (2) and (3). Thus

  $$\begin{aligned} R_{(\sigma_1,\sigma_2,\sigma_0);s}\left(X_{1:n};Y_{1:m}\right) &= \sum_{a\in A} R_{(\sigma_1(a),\sigma_2(a),\sigma_0(a));as}\left(X_{1:n};Y_{1:m}\right) \\ &= \sum_{a\in A} \max_{\bar\sigma_1,\bar\sigma_2,\bar\sigma_0} R_{(\sigma_1,\sigma_2,\sigma_0);as}\left(X_{1:n};Y_{1:m}\right) \\ &= \sum_{a\in A} V_{as}\left(X_{1:n};Y_{1:m}\right) \end{aligned}$$

  by the induction hypothese.

To conclude the proof, it is enough to be reminded that, by definition,

$$V_s\left(X_{1:n};Y_{1:m}\right) = \max\Bigg\{R_{\{\epsilon\};s}\left(X_{1:n};Y_{1:m}\right), \\ V_s\left(X_{1:n}\right)+V_s\left(Y_{1:m}\right), \quad \sum_{a\in A} V_{as}\left(X_{1:n};Y_{1:m}\right)\Bigg\}.$$

Obviously the computational complexity of this procedure is proportional to the number of candidate nodes s, which is equal to the number of distinct subsequences of $X_{1:n}$ and $Y_{1:m}$, and hence quadratic in n and m. However, if necessary, it is possible to obtain a linear complexity algorithm by using compact suffix trees, as explained in Garivier (2006).

5 Simulation study

In this section, we experimentally show the value of joint estimation when the two sources X and Y share some contexts. We compare the results obtained by the BIC joint-estimator described above with the following direct approach. First, we estimate τ_X using the standard BIC tree estimate

$\hat{\tau}_X = \hat{\tau}_X(X_{1:n})$, and we independently estimate τ_Y using $\hat{\tau}_Y = \hat{\tau}_Y(Y_{1:m})$. Then, for all contexts s that are present both in $\hat{\tau}_X$ and in $\hat{\tau}_Y$, we compute the chi-squared distance of the conditional empirical distributions: if this distance is smaller than a given threshold, we decide that s is a shared context. The value of the threshold was chosen in order to maximize the frequency of correct estimation.

5.1 A particularly favorable example

First consider the following case:

- X and Y are $\{1,2\}$-valued context-tree sources;
- $\mathbb{Q}_X$ is defined by the conditional distributions $\mathbb{Q}_X(X_0 = 1|X_{-1} = 1) = 1/3, \mathbb{Q}_X(X_0 = 1|X_{-2:-1} = 12) = 1/3, \mathbb{Q}_X(X_0 = 1|X_{-2:-1} = 22) = 2/3$;
- $\mathbb{Q}_Y$ is defined by the conditional distributions $\mathbb{Q}_Y(Y_0 = 1|Y_{-1} = 1) = 3/4, \mathbb{Q}_Y(Y = 0 = 1|Y_{-2:-1} = 12) = 1/3, \mathbb{Q}_Y(Y_0 = 1|Y_{-2:-1} = 22) = 2/3$;
- the estimates are computed from $X_{1:n}$ and $Y_{1:m}$ with $n = 500$ and $m = 1000$;
- the probability of correctly identifying the tree by each method is estimated by a Monte-Carlo procedure with 1000 replications (margin of error $\approx 1.5\%$).

In that example, we hence have $\sigma_0 = \{12, 22\}$, $\sigma_1 = \{1\}$ and $\sigma_2 = \{1\}$. We compare our joint estimation procedure with separate estimation using the following criteria:

- the probability of correctly identifying τ_X (resp. τ_Y);
- the probability of correctly identifying simultaneously τ_X and τ_Y;
- the probability of correctly identifying $\sigma_0, \sigma_1, \sigma_2$;
- the Kullback-Leibler divergence rates $\mathrm{KL}(\mathbb{Q}_Z|\hat{\mathbb{Q}}_Z)$ between the stationary processes $\mathbb{Q}_Z$ and $\hat{\mathbb{Q}}_Z$ for $Z \in \{X, Y\}$, which are computated by using the fact that both X and Y are Markov chains of finite order.

The results are summarized in Figure 1. It appears that the joint estimation approach has a significant advantage over separate estimation on all the criteria considered here, with one restriction: in some cases, the estimation of either τ_X or τ_Y can be deteriorated, while the other is

	τ_X	τ_Y	τ_X and τ_Y	σ_0	σ_1	σ_2	KL_X	KL_Y
sep. est.	51%	44%	22%	20%	31%	31%	$6.7\ 10^{-3}$	$5.7\ 10^{-3}$
joint est.	80%	78%	76%	77%	90%	90%	$3.2\ 10^{-3}$	$2.3\ 10^{-3}$

Figure 1: Comparative performance of separate and joint estimation in a favorable case (probabilities of correct estimation). KL_X and KL_Y denote $\mathrm{KL}(\mathbb{Q}_X|\hat{\mathbb{Q}}_X)$ and $\mathrm{KL}(\mathbb{Q}_Y|\hat{\mathbb{Q}}_Y)$, respectively.

(more significantly) improved. In all cases, the probability of correctly estimating both τ_X and τ_Y at the same time is increased.

5.2 A less favorable example

On the other hand, when X and Y share no (or few) contexts, then the joint estimation procedure can obviously only deteriorate the separate estimates by introducing some confusion between similar, but distinct conditional distributions of X and Y. An example of such a case is the following:

- X and Y are $\{1, 2\}$-valued context-tree sources;
- $\mathbb{Q}_X$ is defined by the conditional distributions $\mathbb{Q}_X(X_0 = 1|X_{-1} = 1) = 1/2, \mathbb{Q}_X(X_0 = 1|X_{-1} = 2) = 2/3$;
- $\mathbb{Q}_Y$ is defined by the conditional distributions $\mathbb{Q}_Y(Y_0 = 1|Y_{-1} = 1) = 1/2, \mathbb{Q}_Y(Y_0 = 1|Y_{-2:-1} = 12) = 3/5, \mathbb{Q}_Y(Y_0 = 1|Y_{-2:-1} = 22) = 3/4$;
- the estimates are computed from $X_{1:n}$ and $Y_{1:m}$ with $n = 1000$ and $m = 1500$;
- the probability of correctly identifying the tree by each method is estimated by a Monte-Carlo procedure with 1000 replications (margin of error $\approx 1.5\%$).

In that example, $\sigma_0 = \{1\}$, $\sigma_1 = \{2\}$ and $\sigma_2 = \{12, 22\}$. The results are summarized in Figure 2. In this case, $\mathbb{Q}_X$ and $\mathbb{Q}_Y$ are quite close, and the joint estimation procedure tends to merge them into a single, common distribution. Thus, the probability of correctly inferring the structure of $\mathbb{Q}_X$ and $\mathbb{Q}_Y$ is significantly deteriorated.

	τ_X	τ_Y	τ_X and τ_Y	σ_0	σ_1	σ_2	KL_X	KL_Y
sep. est.	97%	89%	86%	84%	84%	82%	$1.0\ 10^{-3}$	$1.3\ 10^{-3}$
joint est.	60%	76%	39%	40%	40%	39%	$1.7\ 10^{-3}$	$2.0\ 10^{-3}$

Figure 2: Comparative performance of separate and joint estimation in the unfavourable case (probabilities of correct estimation). KL_X and KL_Y denote $\mathrm{KL}(\mathbb{Q}_X|\hat{\mathbb{Q}}_X)$ and $\mathrm{KL}(\mathbb{Q}_Y|\hat{\mathbb{Q}}_Y)$, respectively..

5.3 Influence of the penalty term

A natural question is whether the performance of joint (or even separate) estimation can be significantly improved by using other choices of penalty functions, especially choices of the form $\mathrm{pen}(n) = \lambda \log(n)$ for some positive λ. The BIC choice $\lambda = 1$ may be improved by using a recent data-driven procedure called *slope heuristic*, see Birgé & Massart (2007). However, in the present case, the attempts to tune the penalty function by using the slope heuristic merely resulted in a confirmation that the BIC choice could not be significantly improved on the examples considered here. In fact, in addition to the difficulty to detect the dimension gap and thus the minimal penalty in our simulations (which could be expected, as the number of models is very large whereas the sample are not huge), the ideal penalty estimator was never observed to be very different from $\lambda = 1$.

5.4 Discussion

The simulation study strongly indicates that the joint estimation procedure has a significantly improved performance when the two sources do share contexts and conditional distributions which appear with a significant probability in the samples. On the other hand, when the sources share no or few contexts, the procedure may introduce some confusion between the estimates, as could be expected.

When the goal is joint estimation, deterioration in the estimation of one of the trees seems to be the price to pay for better estimating the other tree, and the net effect is positive.

The predictive power of the estimated model is reflected by a measure of discrepancy between the true law of the process and the law of the estimated distribution. We chose to consider Kullback-Leibler divergence, as it is naturally associated to logarithmic prediction loss in information

theory. As expected, a significant improvement is observed for the joint estimator in presence of shared contexts.

References

Birgé, L. & Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138**, 33–73.

Bühlmann, P. & Wyner, A. J. (1999). Variable length Markov chains. *Ann. Statist.* **27**, 480–513.

Csiszár, I. & Talata, Z. (2006). Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inform. Theory* **52**, 1007–1016.

Galves, A., Galves, C., Garcia, N. & Leonardi, F. (2009). Context tree selection and linguistic rhythm retrieval from written texts. *ArXiv:0902.3619* , 1–25.

Garivier, A. (2006). Consistency of the unlimited BIC context tree estimator. *IEEE Trans. Inform. Theory* **52**, 4630–4635.

Garivier, A. & Leonardi, F. (2011). Context tree selection: A unifying view. *Stochastic Processes and their Applications* **121**, 2488–2506.

Gassiat, E. (2010). Codage universel et sélection de modèles emboîtés.

Krichevsky, R. & Trofimov, V. (1981). The performance of universal encoding. *IEEE Trans. Inform. Theory* **27**, 199–207.

Neveu, J. (1972). *Martingales à temps discret.* Masson.

Rissanen, J. (1983). A universal data compression system. *IEEE Trans. Inform. Theory* **29**, 656–664.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

Willems, F., Shtarkov, Y. & Tjalkens, T. (1995). The context-tree weighting method: Basic properties. *IEEE Trans. Inf. Theory* **41**, 653–664.

Addresses

Antonio Galves
Universidade de São Paulo
Instituto de Matemtica e Estatística - USP
Rua do Matão, 1010
CEP 05508-900- São Paulo
Brasil
galves@usp.br

Aurélien Garivier
CNRS UMR 5141
Telecom-ParisTech
LTCI Département TSI
site Dareau
46 rue Barrault 75634 Paris cedex 13
France
garivier@telecom-paristech.fr

Elisabeth Gassiat
Laboratoire de Mathématiques, CNRS UMR 8628
Equipe de Probabilités, Statistique et Modélisation
Université Paris-Sud
Bâtiment 425
91405 Orsay Cedex
France
elisabeth.gassiat@math.u-psud.fr

Appendix

A Technical Lemma

Let $\mathbb{P}_U$ denote the probability distribution of the memoryless source with uniform marginal distribution on A. For a context tree τ and a string $z_{1:k} \in A^k$ denote by $S_\tau(\omega, z_{1:k})$ the concatenation of the symbols that are not in context s for any $s \in \tau$, that is $S_\tau(\omega, z_{1:k}) = \bigodot_{i \in I(z_{1:k}, \tau)} z_i$. Then the Krichevsky-Trofimov Krichevsky & Trofimov (1981) probability distribution is defined as

$$\mathbb{KT}_{(\sigma_0,\sigma_1,\sigma_2)}(x_{1:n}; y_{1:m}) = \mathbb{P}_U\left(S_{\sigma_1\cup\sigma_0}(\omega; x_{1:n})\right)\mathbb{P}_U\left(S_{\sigma_2\cup\sigma_0}(\omega; y_{1:m})\right)$$
$$\prod_{s\in\sigma_0}\mathbb{KT}\left(S(s; x_{1:n}; y_{1:m})\right)\prod_{s\in\sigma_1}\mathbb{KT}\left(S(s; x_{1:n})\right)\prod_{s\in\sigma_2}\mathbb{KT}\left(S(s; y_{1:m})\right), \tag{9}$$

where

$$\mathbb{KT}\left(S(s;x_{1:n};y_{1:m})\right)=\frac{\Gamma\left(\frac{|A|}{2}\right)\prod_{a\in A}\Gamma\left(N_{n,x}\left(s,a\right)+N_{n,y}\left(s,a\right)+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{|A|}\Gamma\left(N_{n,x}\left(s\right)+N_{n,y}\left(s\right)+\frac{|A|}{2}\right)},$$

$$\mathbb{KT}\left(S(s;x_{1:n})\right)=\frac{\Gamma\left(\frac{|A|}{2}\right)\prod_{a\in A}\Gamma\left(N_{n,x}\left(s,a\right)+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{|A|}\Gamma\left(N_{n,x}\left(s\right)+\frac{|A|}{2}\right)},$$

$$\mathbb{KT}\left(S(s;y_{1:m})\right)=\frac{\Gamma\left(\frac{|A|}{2}\right)\prod_{a\in A}\Gamma\left(N_{n,y}\left(s,a\right)+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{|A|}\Gamma\left(N_{n,y}\left(s\right)+\frac{|A|}{2}\right)}.$$

Recall that for any tree σ, $D\left(\sigma\right)$ is its depth :

$$D\left(\sigma\right)=\max\left\{|s|\ :\ s\in\sigma\right\}.$$

Following Willems Willems *et al.* (1995) (see also Gassiat (2010), and references therein), Jensen's inequality leads to the following result:

Lemma 1 *For any $x_{1:n}$ and any $y_{1:m}$,*

$$\begin{aligned}-\log\mathbb{KT}_{(\sigma_0,\sigma_1,\sigma_2)}\left(x_{1:n};y_{1:m}\right)&\leq-\ell_{n,m}\left(\sigma_0,\sigma_1,\sigma_2\right)\\&+\left[D\left(\sigma_0\cup\sigma_1\right)+D\left(\sigma_0\cup\sigma_2\right)+|\sigma_0|+|\sigma_1|+|\sigma_2|\right]\log|A|\\&+\frac{|A|-1}{2}\left\{|\sigma_0|\log\left(\frac{n+m}{|\sigma_0|}\right)+|\sigma_1|\log\left(\frac{n}{|\sigma_1|}\right)+|\sigma_2|\log\left(\frac{m}{|\sigma_2|}\right)\right\}\end{aligned}$$

B Proof of Theorem 1

The proof is divided into four parts.

1. We first prove that eventually almost surely, $|\widehat{\sigma}_0|\leq k_n$ and $|\widehat{\sigma}_1|\leq k_n$ and $|\widehat{\sigma}_2|\leq k_n$ with
$$k_n=\frac{\log n}{\log\log\log n}.$$
For any $(\sigma_0,\sigma_1,\sigma_2)$ satisfying (1), (2) and (3), define $B_{(\sigma_0,\sigma_1,\sigma_2)}$ as the set of $(x_{1:n},y_{1:m})$ in A^{n+m} such that
$$(X_{1:n},Y_{1:m})=(x_{1:n},y_{1:m})\Leftrightarrow(\widehat{\sigma}_0,\widehat{\sigma}_1,\widehat{\sigma}_2)=(\sigma_0,\sigma_1,\sigma_2),$$
so that
$$\begin{aligned}&\mathbb{Q}^*\left((\widehat{\sigma}_0,\widehat{\sigma}_1,\widehat{\sigma}_2)=(\sigma_0,\sigma_1,\sigma_2)\right)\\&\quad=\sum_{(x_{1:n},y_{1:m})\in B_{(\sigma_0,\sigma_1,\sigma_2)}}\mathbb{Q}^*\left((X_{1:n},Y_{1:m})=(x_{1:n},y_{1:m})\right).\end{aligned}$$

If $(X_{1:n}, Y_{1:m}) \in B_{(\sigma_0,\sigma_1,\sigma_2)}$, then

$$\begin{aligned}&\ell_{n,m}(\sigma_0,\sigma_1,\sigma_2) - \frac{(|A|-1)}{2}(|\sigma_0|\mathrm{pen}(n+m) + |\sigma_1|\mathrm{pen}(n) + |\sigma_2|\mathrm{pen}(m))\\ &\geq \ell_{n,m}(\sigma_0^*,\sigma_1^*,\sigma_2^*) - \frac{(|A|-1)}{2}(|\sigma_0^*|\mathrm{pen}(n+m) + |\sigma_1^*|\mathrm{pen}(n) + |\sigma_2^*|\mathrm{pen}(m)),\end{aligned}$$

and using Lemma 1, if $(x_{1:n}, y_{1:m}) \in B_{(\sigma_0,\sigma_1,\sigma_2)}$, then

$$\begin{aligned}\mathbb{Q}^*((x_{1:n}, y_{1:m})) &\leq 2^{\ell_{n,m}\left(\sigma_0^*,\sigma_1^*,\sigma_2^*\right)}\\ &\leq 2^{\ell_{n,m}(\sigma_0,\sigma_1,\sigma_2) + \frac{(|A|-1)}{2}\left(\left(|\sigma_0^*|-t_0\right)\mathrm{pen}(n+m) + \left(|\sigma_1^*|-t_1\right)\mathrm{pen}(n) + \left(|\sigma_2^*|-t_2\right)\mathrm{pen}(m)\right)}\\ &\leq \mathbb{KT}_{(\sigma_0,\sigma_1,\sigma_2)}(x_{1:n}; y_{1:m})\, 2^{H(n,m,t_0,t_1,t_2)}\end{aligned}$$

with $t_i = |\sigma_i|$, $i = 0, 1, 2$, and

$$\begin{aligned}&H\big(n,m,t_0,t_1,t_2\big) =\\ &\quad \frac{|A|-1}{2}\left\{t_0 \log\left(\frac{n+m}{t_0}\right) + t_1 \log\left(\frac{n}{t_1}\right) + t_2 \log\left(\frac{m}{t_2}\right)\right\}\\ &\quad + \frac{(|A|-1)}{2}\left((|\sigma_0^*| - t_0)\mathrm{pen}(n+m) + (|\sigma_1^*| - t_1)\mathrm{pen}(n) + (|\sigma_2^*| - t_2)\mathrm{pen}(m)\right)\\ &\quad + [3t_0 + 2t_1 + 2t_2]\log|A|\\ &\quad = \frac{|A|-1}{2}\Big\{ - t_0\log t_0 - t_1\log t_1 - t_2\log t_2 + |\sigma_0^*|\log(n+m) +\\ &\quad |\sigma_1^*|\log(n) + |\sigma_2^*|\log(m)\Big\} + [3t_0 + 2t_1 + 2t_2]\log|A|\end{aligned}$$

using $\mathrm{pen}(\cdot) = \log(\cdot)$ and using that for a complete tree σ, $D(\sigma) \leq |\sigma|$.
Thus,

$$\mathbb{Q}^*\left((\widehat{\sigma}_0, \widehat{\sigma}_1, \widehat{\sigma}_2) = (\sigma_0,\sigma_1,\sigma_2)\right) \leq 2^{H(n,m,t_0,t_1,t_2)},$$

and

$$\begin{aligned}&\mathbb{Q}^*\left(|\widehat{\sigma}_0| \geq k_n \text{ or } |\widehat{\sigma}_1| \geq k_n \text{ or } |\widehat{\sigma}_2| \geq k_n\right)\\ &\qquad\leq \sum_{t_0=k_n+1}^{n\vee m}\sum_{t_1,t_2=0}^{n\vee m} F(t_0,t_1,t_2)\, 2^{H(n,m,t_0,t_1,t_2)}\\ &\qquad+ \sum_{t_1=k_n+1}^{n\vee m}\sum_{t_0,t_2=0}^{n\vee m} F(t_0,t_1,t_2)\, 2^{H(n,m,t_0,t_1,t_2)}\\ &\qquad\qquad+ \sum_{t_2=k_n+1}^{n\vee m}\sum_{t_0,t_1=0}^{n\vee m} F(t_0,t_1,t_2)\, 2^{H(n,m,t_0,t_1,t_2)}\end{aligned}$$

where $F(t_0,t_1,t_2)$ is the number of $(\sigma_0,\sigma_1,\sigma_2)$ satisfying (1), (2) and (3) and such that $|\sigma_0| = t_0$, $|\sigma_1| = t_1$, and $|\sigma_2| = t_2$.

But the number of complete trees with t elements is upper bounded by 16^t, see Garivier (2006), so that, denoting by $\binom{b}{a} \leq 2^b$ the binomial coefficient, one has

$$\begin{aligned} F\left(t_0, t_1, t_2\right) &\leq \binom{t_0+t_1}{t_0} 16^{t_0+t_1} \binom{t_0+t_2}{t_0} 16^{t_0+t_2} \\ &\leq 16^{4t_0+2t_1+2t_2}. \end{aligned}$$

Using the fact that for any constant a, $-t\log t + at$ is bounded on $\mathbb{R}^+$, and using (6) one gets that for some constants C_1, C_2 and C_3,

$$\mathbb{Q}^*\left(|\widehat{\sigma}_0| \geq k_n \text{ or } |\widehat{\sigma}_1| \geq k_n \text{ or } |\widehat{\sigma}_2| \geq k_n\right) \leq C_1 2^{-C_2 k_n \log k_n + C_3 \log n}.$$

But

$$\lim_{n\to+\infty} \frac{k_n \log k_n}{\log n} = +\infty$$

so that one gets that for another constant C,

$$\mathbb{Q}^*\left(|\widehat{\sigma}_0| \geq k_n \text{ or } |\widehat{\sigma}_1| \geq k_n \text{ or } |\widehat{\sigma}_2| \geq k_n\right) \leq \frac{C}{n^2}$$

and using Borel-Cantelli's Lemma, we obtain that $\mathbb{Q}^*$-eventually almost surely, $|\widehat{\sigma}_0| \leq k_n$ and $|\widehat{\sigma}_1| \leq k_n$ and $|\widehat{\sigma}_2| \leq k_n$.

2. We prove that $\mathbb{Q}^*$-eventually almost surely, no context is overestimated.

   It is sufficient to prove that, $\mathbb{Q}^*$-almost surely, if $(\sigma_0, \sigma_1, \sigma_2)$ satisfy (1), (2) and (3) and are such that for some i, σ_i contains some string that has a proper suffix in σ_i^*, there exists $(\bar{\sigma}_0, \bar{\sigma}_1, \bar{\sigma}_2)$ satisfying (1), (2) and (3) and such that, eventually, $C_{n,m}(\bar{\sigma}_0, \bar{\sigma}_1, \bar{\sigma}_2) > C_{n,m}(\sigma_0, \sigma_1, \sigma_2)$, so that $(\widehat{\sigma}_0, \widehat{\sigma}_1, \widehat{\sigma}_2) \neq (\sigma_0, \sigma_1, \sigma_2)$.
   Consider first the case where σ_0^* is overestimated. Let $(\sigma_0, \sigma_1, \sigma_2)$ satisfy (1), (2) and (3) and be such that σ_0 contains some string that has a proper suffix in σ_0^*. Let $s = av$, $a \in A$, be the longest such string, and let $u \in \sigma_0^*$ be the corresponding suffix of v. For $i \in \{0, 1, 2\}$, let $S_i = A^+ v \cap \sigma_i$ and define

$$\bar{\sigma}_0 = (\sigma_0 \backslash S_0) \cup \{v\}\ , \qquad \bar{\sigma}_1 = (\sigma_1 \backslash S_1)\ , \qquad \bar{\sigma}_2 = (\sigma_2 \backslash S_2)\ .$$

Then

$$
\begin{aligned}
& C_{n,m}(\bar{\sigma}_0, \bar{\sigma}_1, \bar{\sigma}_2) - C_{n,m}(\sigma_0, \sigma_1, \sigma_2) \\
= \quad & \sum_{b\in A} \left[N_{n,X}(v,b) + N_{m,Y}(v,b)\right] \log\left(\frac{N_{n,X}(v,b) + N_{m,Y}(v,b)}{N_{n,X}(v) + N_{m,Y}(v)}\right) \\
& \quad - \frac{|A|-1}{2}\log(n+m) \\
& - \sum_{w\in S_0} \Bigg\{ \sum_{b\in A} \left[N_{n,X}(w,b) + N_{m,Y}(w,b)\right] \log\left(\frac{N_{n,X}(w,b) + N_{m,Y}(w,b)}{N_{n,X}(w) + N_{m,Y}(w)}\right) \\
& \quad - \frac{|A|-1}{2}\log(n+m) \Bigg\} \\
& - \sum_{w\in S_1} \left\{ \sum_{b\in A} N_{n,X}(w,b) \log\left(\frac{N_{n,X}(w,b)}{N_{n,X}(w)}\right) - \frac{|A|-1}{2}\log(n) \right\} \\
& - \sum_{w\in S_2} \left\{ \sum_{b\in A} N_{m,Y}(w,b) \log\left(\frac{N_{m,Y}(w,b)}{N_{m,Y}(w)}\right) - \frac{|A|-1}{2}\log(m) \right\}
\end{aligned}
$$

By definition of the maximum likelihood, the above expression is lower-bounded by:

$$
\begin{aligned}
& C_{n,m}(\bar{\sigma}_0, \bar{\sigma}_1, \bar{\sigma}_2) - C_{n,m}(\sigma_0, \sigma_1, \sigma_2) \\
\geq \quad & \sum_{b\in A} \left[N_{n,X}(v,b) + N_{m,Y}(v,b)\right] \log\left(Q_X^*(b|v)\right) - \frac{|A|-1}{2}\log(n+m) \\
& - \sum_{w\in S_0} \Bigg\{ \sum_{b\in A} \left[N_{n,X}(w,b) + N_{m,Y}(w,b)\right] \log\left(\widehat{Q}_{XY}(b|w)\right) \\
& \quad - \frac{|A|-1}{2}\log(n+m) \Bigg\} \\
& - \sum_{w\in S_1} \left\{ \sum_{b\in A} N_{n,X}(w,b) \log\left(\widehat{Q}_X(b|w)\right) - \frac{|A|-1}{2}\log(n) \right\} \\
& - \sum_{w\in S_2} \left\{ \sum_{b\in A} N_{m,Y}(w,b) \log\left(\widehat{Q}_Y(b|w)\right) - \frac{|A|-1}{2}\log(m) \right\}
\end{aligned}
$$

Notice that

$$Q_X^*(\cdot|v) = Q_Y^*(\cdot|v) = Q_X^*(\cdot|w)$$

for any $w \in S_0 \cup S_1 \cup S_2$.

It follows from part 1 of the proof that we only need to consider trees σ_i such that $|\sigma_i| = o(\log n)$. Notice also that since $D(\sigma_i) = o(\log n)$, for any $b \in A$,

$$N_{n,X}(v,b) = \sum_{w\in S_0\cup S_1} N_{n,X}(w,b) + o(\log n),$$

$$N_{m,Y}(v,b) = \sum_{w\in S_0\cup S_2} N_{m,Y}(w,b) + o(\log n).$$

Let $\mathrm{KL}\left(q_1|q_2\right) = \sum_{a\in A} q_1(a)\log\frac{q_1(a)}{q_2(a)}$ denotes the Kullback-Leibler divergence between two probability measures q_1 and q_2 on A, with the convention that $0\log(0/x) = 0$ for $x \geq 0$ and $x\log(x/0) = +\infty$ for $x > 0$. Since the minimum of all positive transition probabilities in $\mathbb{Q}^*$ is positive, one gets

$$\begin{aligned}
&C_{n,m}(\bar\sigma_0,\bar\sigma_1,\bar\sigma_2) - C_{n,m}(\sigma_0,\sigma_1,\sigma_2)\\
&\geq \sum_{w\in S_0}\sum_{b\in A}\left[N_{n,X}(w,b)+N_{m,Y}(w,b)\right]\log\left(\frac{Q^*_X(b|w)}{\widehat{Q}_{XY}(b|w)}\right)\\
&\qquad + \left(|S_0|-1\right)\frac{|A|-1}{2}\log(n+m)\\
&\quad + \sum_{w\in S_1}\sum_{b\in A} N_{n,X}(w,b)\log\left(\frac{Q^*_X(b|w)}{\widehat{Q}_X(b|w)}\right) + |S_1|\frac{|A|-1}{2}\log(n)\\
&\quad + \sum_{w\in S_2}\sum_{b\in A} N_{m,Y}(w,b)\log\left(\frac{Q^*_Y(b|w)}{\widehat{Q}_Y(b|w)}\right) + |S_2|\frac{|A|-1}{2}\log(m)\\
&\quad + o(\log n)\\
&= -\sum_{w\in S_0}\left[N_{n,X}(w)+N_{m,Y}(w)\right]\mathrm{KL}\left(\widehat{Q}_{XY}(\cdot|w)\,|Q^*_X(\cdot|w)\right)\\
&\qquad + \left(|S_0|-1\right)\frac{|A|-1}{2}\log(n+m)\\
&\quad - \sum_{w\in S_1} N_{n,X}(w)\,\mathrm{KL}\left(\widehat{Q}_X(\cdot|w)\,|Q^*_X(\cdot|w)\right) + |S_1|\frac{|A|-1}{2}\log(n)\\
&\quad - \sum_{w\in S_2} N_{m,Y}(w)\,\mathrm{KL}\left(\widehat{Q}_Y(\cdot|w)\,|Q^*_Y(\cdot|w)\right) + |S_2|\frac{|A|-1}{2}\log(m)\\
&\quad + o(\log n).
\end{aligned}$$

According to typicality Lemma 6.2 of Csiszár & Talata (2006), for all $\delta > 0$, for all w such that $N_{n,X}(w) \geq 1$ and for all $b \in A$ it holds that, $\mathbb{Q}^*$-eventually almost surely,

$$\left|\widehat{Q}_X(b|w) - Q^*_X(b|w)\right| \leq \sqrt{\frac{\delta\log(n)}{N_{n,X}(w)}}\,.$$

Besides, Lemma 6.3 of Csiszár & Talata (2006) states that

$$\mathrm{KL}\left(\widehat{Q}_X(\cdot|w)\,|Q^*_X(\cdot|w)\right) \leq \sum_{b\in A}\frac{\left(\widehat{Q}_X(b|w) - Q^*_X(b|w)\right)^2}{Q^*_X(b|w)}\,.$$

Handling similarly the terms involving Q^*_Y and Q^*_{XY}, and denoting $q^*_{min} > 0$ the minimum of all positive transition probabilities in $\mathbb{Q}^*$,

we obtain that for any $\delta > 0$, $\mathbb{Q}^*$-eventually almost surely for all possible $(\sigma_0, \sigma_1, \sigma_2)$:

$$\begin{aligned}
&C_{n,m}(\bar\sigma_0, \bar\sigma_1, \bar\sigma_2) - C_{n,m}(\sigma_0, \sigma_1, \sigma_2) \geq \\
&\quad -\frac{\delta|A|}{q^*_{min}}|S_0| \log(n+m) + (|S_0| - 1)\,\frac{|A|-1}{2} \log(n+m) \\
&\qquad -\frac{\delta|A|}{q^*_{min}}|S_1| \log(n) + |S_1|\frac{|A|-1}{2} \log(n) \\
&\qquad\quad -\frac{\delta|A|}{q^*_{min}}|S_2| \log(m) + |S_2|\frac{|A|-1}{2} \log(m)
\end{aligned}$$

which is positive, for all possible $(\sigma_0, \sigma_1, \sigma_2)$, $\mathbb{Q}^*$-eventually almost surely. This follows from the fact that we consider complete context trees, and therefore $|S_0| \geq 1$, $|S_0| + |S_1| \geq |A|$ and $|S_0| + |S_2| \geq |A|$. Consider now the case where σ^*_i, $i = 1$ or $i = 2$ is overestimated. Let $(\sigma_0, \sigma_1, \sigma_2)$ satisfy (1), (2) and (3) and be such that σ_i contains some string that has a proper suffix in σ^*_i. Let $s = av$, $a \in A$, be the longest such string, and let $u \in \sigma^*_i$ be the corresponding suffix of v. For $i = 0, 1, 2$, let again, $S_i = A^+v \cap \sigma_i$. Then, either $S_0 = \emptyset$, and the problem boils down the the overestimation of a single tree: the consistency result of Csiszár & Talata (2006) applies and shows that denoting

$$\bar\sigma_i = (\sigma_1 \backslash S_i) \cup \{v\} \ , \quad \bar\sigma_j = \sigma_j, j \neq i \ ,$$

we have $C_{n,m}(\bar\sigma_0, \bar\sigma_1, \bar\sigma_2) > C_{n,m}(\sigma_0, \sigma_1, \sigma_2)$ $\mathbb{Q}^*$-eventually almost surely. Or σ^*_0 has also been overestimated, so that one may apply the previous proof.

3. Consider now the underestimation case. If σ_0 has been underestimated, there exists $s \in \sigma_0$ which is a proper suffix of $s_0 \in \sigma^*_0$. For $i = 0, 1, 2$, let $S_i = A^+s \cap \sigma^*_i$, and define

$$\bar\sigma_0 = (\sigma_0 \backslash \{s\}) \cup S_0 \ , \quad \bar\sigma_1 = \sigma_1 \cup S_1 \ , \quad \bar\sigma_2 = \sigma_2 \cup S_2.$$

Then

$$
\begin{aligned}
&C_{n,m}(\bar{\sigma}_0, \bar{\sigma}_1, \bar{\sigma}_2) - C_{n,m}(\sigma_0, \sigma_1, \sigma_2) \\
&\quad = \sum_{w \in S_0} \left\{ \sum_{b \in A} \left[N_{n,X}\left(w,b\right) + N_{m,Y}\left(w,b\right)\right] \log\left(\frac{N_{n,X}\left(w,b\right) + N_{m,Y}\left(w,b\right)}{N_{n,X}\left(w\right) + N_{m,Y}\left(w\right)} \right) \right. \\
&\qquad\qquad \left. - \frac{|A|-1}{2} \log\left(n+m\right) \right\} \\
&\qquad + \sum_{w \in S_1} \left\{ \sum_{b \in A} N_{n,X}\left(w,b\right) \log\left(\frac{N_{n,X}\left(w,b\right)}{N_{n,X}\left(w\right)} \right) - \frac{|A|-1}{2} \log\left(n\right) \right\} \\
&\qquad + \sum_{w \in S_2} \left\{ \sum_{b \in A} N_{m,Y}\left(w,b\right) \log\left(\frac{N_{m,Y}\left(w,b\right)}{N_{m,Y}\left(w\right)} \right) - \frac{|A|-1}{2} \log\left(m\right) \right\} \\
&\qquad - \sum_{b \in A} \left[N_{n,X}\left(s,b\right) + N_{m,Y}\left(s,b\right)\right] \log\left(\frac{N_{n,X}\left(s,b\right) + N_{m,Y}\left(s,b\right)}{N_{n,X}\left(s\right) + N_{m,Y}\left(v\right)} \right) \\
&\qquad\qquad + \frac{|A|-1}{2} \log\left(n+m\right)
\end{aligned}
$$

Notice that for any string u, for any $b \in A$, $\frac{1}{n} N_{n,X}\left(u,b\right)$ and $\frac{1}{n} N_{n,X}\left(u\right)$ converge $\mathbb{Q}^*$ almost surely to $Q_X^*\left(ub\right)$ and $Q_X^*\left(u\right)$ respectively, and $\frac{1}{n} N_{m,Y}\left(u,b\right)$ and $\frac{1}{n} N_{m,Y}\left(u\right)$ converge $\mathbb{Q}^*$ almost surely to $\frac{1}{c} Q_Y^*\left(ub\right)$ and $\frac{1}{c} Q_Y^*\left(u\right)$, respectively.

Thus, $\mathbb{Q}^*$ almost surely,

$$
\begin{aligned}
&C_{n,m}(\bar{\sigma}_0, \bar{\sigma}_1, \bar{\sigma}_2) - C_{n,m}(\sigma_0, \sigma_1, \sigma_2) = -O\left(\log n\right) \\
&+ n \sum_{w \in S_0} \sum_{b \in A} \left[Q_X^*\left(wb\right) + \frac{1}{c} Q_Y^*\left(wb\right) \right] \log\left(\frac{Q_X^*\left(wb\right) + \frac{1}{c} Q_Y^*\left(wb\right)}{Q_X^*\left(w\right) + \frac{1}{c} Q_Y^*\left(w\right)} \right) \\
&+ n \sum_{w \in S_1} \sum_{b \in A} Q_X^*\left(wb\right) \log\left(\frac{Q_X^*\left(wb\right)}{Q_X^*\left(w\right)} \right) \\
&+ n \sum_{w \in S_2} \sum_{b \in A} \frac{1}{c} Q_Y^*\left(wb\right) \log\left(\frac{Q_Y^*\left(wb\right)}{Q_Y^*\left(w\right)} \right) \\
&- n \sum_{b \in A} \left[Q_X^*\left(sb\right) + \frac{1}{c} Q_Y^*\left(sb\right) \right] \log\left(\frac{Q_X^*\left(sb\right) + \frac{1}{c} Q_Y^*\left(sb\right)}{Q_X^*\left(s\right) + \frac{1}{c} Q_Y^*\left(s\right)} \right) + o\left(n\right) \\
&= -O\left(\log n\right) + o\left(n\right) + n \sum_{w \in S_0 \cup S_1} \sum_{b \in A} Q_X^*\left(wb\right) \log\left(\frac{Q_X^*\left(wb\right)}{Q_X^*\left(w\right)} \right) \\
&+ n \sum_{w \in S_0 \cup S_2} \sum_{b \in A} \frac{1}{c} Q_Y^*\left(wb\right) \log\left(\frac{Q_Y^*\left(wb\right)}{Q_Y^*\left(w\right)} \right) \\
&- n \sum_{b \in A} \left[Q_X^*\left(sb\right) + \frac{1}{c} Q_Y^*\left(sb\right) \right] \log\left(\frac{Q_X^*\left(sb\right) + \frac{1}{c} Q_Y^*\left(sb\right)}{Q_X^*\left(s\right) + \frac{1}{c} Q_Y^*\left(s\right)} \right)
\end{aligned}
$$

because for $w \in S_0$, $Q_X^*(wb) = Q_Y^*(wb)$. Since

$$\sum_{w \in S_0 \cup S_1} Q_X^*(w) = Q_X^*(s),$$

for any $b \in A$, Jensen's inequality implies that

$$\sum_{w \in S_0 \cup S_1} Q_X^*(wb) \log\left(\frac{Q_X^*(wb)}{Q_X^*(w)}\right) \geq Q_X^*(sb) \log\left(\frac{Q_X^*(sb)}{Q_X^*(s)}\right),$$

and the inequality is strict for at least one $b \in A$, for otherwise, s would be a context for $\mathbb{Q}_X^*$. Similarly for any $b \in A$,

$$\sum_{w \in S_0 \cup S_2} Q_Y^*(wb) \log\left(\frac{Q_Y^*(wb)}{Q_Y^*(w)}\right) \geq Q_Y^*(sb) \log\left(\frac{Q_Y^*(sb)}{Q_Y^*(s)}\right).$$

Using the concavity of the entropy function

$$\begin{aligned}
&\sum_{b \in A} Q_X^*(sb) \log\left(\frac{Q_X^*(sb)}{Q_X^*(s)}\right) + \frac{1}{c}\sum_{b \in A} Q_Y^*(sb) \log\left(\frac{Q_Y^*(sb)}{Q_Y^*(s)}\right) \\
&\geq \sum_{b \in A}\left(Q_X^*(sb) + \frac{1}{c}Q_Y^*(sb)\right) \log\left(\frac{Q_X^*(sb) + \frac{1}{c}Q_Y^*(sb)}{Q_X^*(s) + \frac{1}{c}Q_Y^*(s)}\right),
\end{aligned}$$

so that there exists $\delta > 0$ such that

$$C_{n,m}(\bar{\sigma}_0, \bar{\sigma}_1, \bar{\sigma}_2) - C_{n,m}(\sigma_0, \sigma_1, \sigma_2) \geq n\delta$$

$\mathbb{Q}^*$-eventually almost surely.

If σ_i, $i = 1$ or $i = 2$ has been underestimated, then the problem boils down to the standard underestimation of a single context tree. Defining (with obvious notation)

$$\bar{\sigma}_i = (\sigma_1 \setminus \{s\}) \cup S_i \cup S_0\,, \quad \bar{\sigma}_j = \sigma_j, j \neq i\,,$$

it is proved in Csiszár & Talata (2006), Section III, that $\mathbb{Q}^*$-eventually almost surely, $C_{n,m}(\bar{\sigma}_0, \bar{\sigma}_1, \bar{\sigma}_2) > C_{n,m}(\sigma_0, \sigma_1, \sigma_2)$.

4. We have thus proved that, for $i = 1$ and $i = 2$, $\widehat{\sigma}_0 \cup \widehat{\sigma}_i = \sigma_0^* \cup \sigma_i^*$, $\mathbb{Q}^*$-eventually almost surely. Let $(\sigma_0, \sigma_1, \sigma_2)$ satisfy (1), (2) and (3) and be such that, for $i = 1$ and $i = 2$, $\sigma_0 \cup \sigma_i = \sigma_0^* \cup \sigma_i^*$. There remains to check that $\mathbb{Q}^*$ almost surely, if there exists a string s such that

   - $s \in \sigma_0$, but $s \in \sigma_1^*$ and $s \in \sigma_2^*$,
   - or $s \in \sigma_1$ and $s \in \sigma_2$, but $s \in \sigma_0^*$,

then $(\widehat{\sigma_0}, \widehat{\sigma_1}, \widehat{\sigma_2}) \neq (\sigma_0, \sigma_1, \sigma_2)$ eventually.
Consider first the case where $s \in \sigma_0$, but $s \in \sigma_1^*$ and $s \in \sigma_2^*$. Define

$$\bar{\sigma}_0 = (\sigma_0 \backslash \{s\}) \ , \quad \bar{\sigma}_1 = \sigma_1 \cup \{s\} \ , \quad \bar{\sigma}_2 = \sigma_2 \cup \{s\} \ .$$

Then

$$\begin{aligned}
&C_{n,m}(\bar{\sigma}_0, \bar{\sigma}_1, \bar{\sigma}_2) - C_{n,m}(\sigma_0, \sigma_1, \sigma_2) = \\
&+ \sum_{b \in A} N_{n,X}\left(s,b\right) \log \left(\frac{N_{n,X}\left(s,b\right)}{N_{n,X}\left(s\right)} \right) \\
&\sum_{b \in A} N_{m,Y}\left(s,b\right) \log \left(\frac{N_{m,Y}\left(s,b\right)}{N_{m,Y}\left(s\right)} \right) \\
&- \sum_{b \in A} \left[N_{n,X}\left(s,b\right) + N_{m,Y}\left(s,b\right) \right] \log \left(\frac{N_{n,X}\left(s,b\right) + N_{m,Y}\left(s,b\right)}{N_{n,X}\left(s\right) + N_{m,Y}\left(s\right)} \right) \\
&+ \frac{|A|-1}{2} \left\{ \log\left(n+m\right) - \log n - \log m \right\} \\
&= n \left\{ \sum_{b \in A} Q_X^*\left(sb\right) \log \left(\frac{Q_X^*\left(sb\right)}{Q_X^*\left(s\right)} \right) + \frac{1}{c} \sum_{b \in A} Q_Y^*\left(sb\right) \log \left(\frac{Q_Y^*\left(sb\right)}{Q_Y^*\left(s\right)} \right) \right. \\
&\left. - \sum_{b \in A} \left(Q_X^*\left(sb\right) + \frac{1}{c} Q_Y^*\left(sb\right) \right) \log \left(\frac{Q_X^*\left(sb\right) + \frac{1}{c} Q_Y^*\left(sb\right)}{Q_X^*\left(s\right) + \frac{1}{c} Q_Y^*\left(s\right)} \right) + o(1) \right\} \\
&- O\left(\log n\right)
\end{aligned}$$

$\mathbb{Q}^*$ almost surely. But the quantity into brackets is positive by the strict concavity of the entropy function, unless for any $b \in A$, $Q_X^*(b|s) = Q_Y^*(b|s)$ which would mean that $s \in \sigma_0^*$.
Consider now the case where $s \in \sigma_1$ and $s \in \sigma_2$, but $s \in \sigma_0^*$. Define

$$\begin{aligned}
\bar{\sigma}_0 &= \sigma_0 \cup \{s\}, \\
\bar{\sigma}_1 &= (\sigma_1 \backslash \{s\}) \,, \\
\bar{\sigma}_2 &= (\sigma_2 \backslash \{s\}) \,.
\end{aligned}$$

$$\begin{aligned}
&C_{n,m}(\bar{\sigma}_0, \bar{\sigma}_1, \bar{\sigma}_2) - C_{n,m}(\sigma_0, \sigma_1, \sigma_2) = \\
&\sum_{b \in A} \left[N_{n,X}\left(s,b\right) + N_{m,Y}\left(s,b\right) \right] \log \left(\frac{N_{n,X}\left(s,b\right) + N_{m,Y}\left(s,b\right)}{N_{n,X}\left(s\right) + N_{m,Y}\left(s\right)} \right) \\
&- \sum_{b \in A} N_{n,X}\left(s,b\right) \log \left(\frac{N_{n,X}\left(s,b\right)}{N_{n,X}\left(s\right)} \right) \\
&- \sum_{b \in A} N_{m,Y}\left(s,b\right) \log \left(\frac{N_{m,Y}\left(s,b\right)}{N_{m,Y}\left(s\right)} \right) \\
&+ \frac{|A|-1}{2} \left\{ \log n + \log m - \log\left(n+m\right) \right\} .
\end{aligned}$$

Using Taylor expansion until second order of $u \log u$, one gets

$$
\begin{aligned}
&C_{n,m}(\bar{\sigma}_0, \bar{\sigma}_1, \bar{\sigma}_2) - C_{n,m}(\sigma_0, \sigma_1, \sigma_2) \\
&= \left\{ \frac{1}{2} \sum_{b \in A} \frac{\left([N_{n,X}(s,b) + N_{m,Y}(s,b)] - [N_{n,X}(s) + N_{m,Y}(s)]\, Q^*_X(b|s) \right)^2}{[N_{n,X}(s) + N_{m,Y}(s)]\, Q^*_X(b|s)} \right. \\
&\quad - \frac{1}{2} \sum_{b \in A} \frac{\left(N_{n,X}(s,b) - N_{n,X}(s)\, Q^*_X(b|s) \right)^2}{N_{n,X}(s)\, Q^*_X(b|s)} \\
&\quad \left. - \frac{1}{2} \sum_{b \in A} \frac{\left(N_{m,Y}(s,b) - N_{m,Y}(s)\, Q^*_Y(b|s) \right)^2}{N_{m,Y}(s)\, Q^*_Y(b|s)} \right\} (1 + o(1)) \\
&\quad + \frac{|A| - 1}{2} \left\{ \log n + \log m - \log(n+m) \right\}.
\end{aligned}
$$

The sequences

$$
\begin{aligned}
&\left(N_{n,X}(s,b) - N_{n,X}(s)\, Q^*_X(b|s) \right)_{n \geq 0}, \\
&\left(N_{m,Y}(s,b) - N_{m,Y}(s)\, Q^*_Y(b|s) \right)_{m \geq 0},
\end{aligned}
$$

are martingales with respect to the natural filtration. Thus, it follows from the law of iterated logarithm for martingales Neveu (1972) that, $\mathbb{Q}^*$ almost surely,

$$
\begin{aligned}
C_{n,m}(\bar{\sigma}_0, \bar{\sigma}_1, \bar{\sigma}_2) - C_{n,m}(\sigma_0, \sigma_1, \sigma_2) &= O(\log \log n) \\
&\quad + \frac{|A| - 1}{2} \left\{ \log n + \log m - \log(n+m) \right\},
\end{aligned}
$$

so that $\mathbb{Q}^*$ almost surely,

$$
C_{n,m}(\bar{\sigma}_0, \bar{\sigma}_1, \bar{\sigma}_2) - C_{n,m}(\sigma_0, \sigma_1, \sigma_2) > 0
$$

eventually. This ends the proof of Theorem 1.

Acknowledgments

This work is part of USP project MaCLinC, “Mathematics, computation, language and the brain”, USP/COFECUB project “Stochastic systems with interactions of variable range” and CNPq project *Rhythmic patterns, prosodic domains and probabilistic modeling in Portuguese Corpora* (grant number 485999/2007-2). This paper was partially supported by CAPES grant AUXPE-PAE-598/2011. A. Galves is partially supported by a CNPq fellowship (grant 305447/2008-4). The authors thank P. Weyer-Brown for his help in English.